NREL
NATIONAL RENEWABLE ENERGY LABORATORY

NREL is a national laboratory of the U.S. Department of Energy, Office of Energy Efficiency & Renewable Energy, operated by the Alliance for Sustainable Energy, LLC.

Automated Analysis of Renewable Energy Datasets ("EE/RE Data Mining")

Brian Bush, Ryan Elmore, Dan Getman, Danny Inman, Eric Kalendra

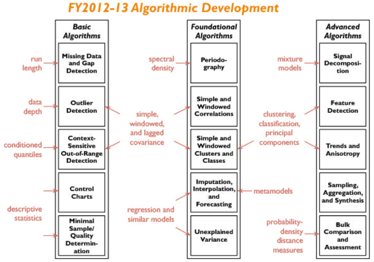
Goals, Plans, Impacts

Goals: To dramatically improve the understanding of EE/RE (energy efficiency and renewable energy) data sets and the depth and efficiency of their analysis through the application of statistical learning methods ("data mining") in the intelligent processing of these often large and messy information sources.

Focus Areas:
- anomaly detection
- data cleansing
- forecasting
- pattern mining
- reduced-complexity models
- automated reasoning

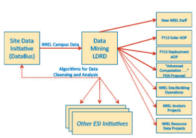
Application Areas:
- time-series data (e.g., campus meter data)
- spatiotemporal data (e.g., resource datasets)
- complex semi-structured data (e.g., incentives databases)

FY2012-13 Algorithmic Development



Impact:
Results from this project have the potential to routinely add value to a wide range of projects across most, if not all, NREL centers.
- The rapid and efficient data mining techniques can significantly lower the costs associated with analysis of data within data-intensive projects and become a standard feature of such projects.
- This project will provide increased leverage to both NREL and EE/RE data sets.
- The addition of such a capability to NREL can be harnessed in marketing sophisticated, complex analysis projects to multiple sponsors, putting NREL another step ahead of competitors.

Concrete results:



- Concepts and approaches from this project have been and are being incorporated into FY2013-14 proposals and work plans.
- Work is being coordinated with other Energy Systems Integration (ESI) initiatives.
- Understanding and results have contributed to the DataBus initiative.
- A data-science postdoc was hired.

Next Steps:
Packaging and documentation of R code:
- Fuzzy autocorrelation-function-based (ACF) clustering of NREL campus data
- Wavelet clustering of NREL campus data
- Anomaly detection, filling missing data, and denoising of irregularly gridded time series
- Detection of spatial and temporal trends in large datasets
- Gaussian process modeling of multivariate time series

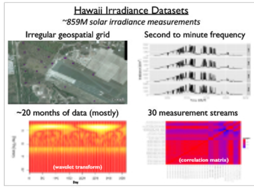
Reports, conference papers, and/or journal papers:
- Diagnostics, clustering, imputation, and forecasting for NREL campus data.
- Decorrelated multiscale covariate model for irregularly gridded time series.
- Gaussian process modeling of multivariate time series.

Document for NREL and EE/RE "big data" lessons learned

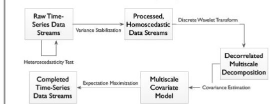
Statistical Modeling of Multivariate Time Series

Goals: Develop a widely reusable minimalistic model (e.g., in the spirit of Occam's Razor), for processing datasets consisting of multiple time series on an irregular spatial grid, for . . .
- filling in (interpolating) missing data
- removing noise or smoothing data for applications new requiring high resolution
- identifying anomalous data points ("outliers") and patterns
- extrapolation and forecasting

Results: We developed two general purpose models (see below) practically meeting these requirements.

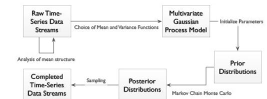


Multiscale Covariate (MC) Model

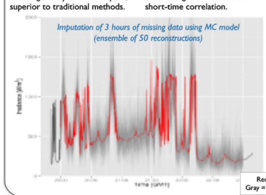


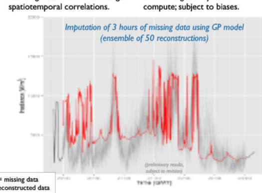
Advantages: very fast to compute; superior to traditional methods.
Disadvantages: misses some of short-time correlation.

Gaussian Process (GP) Model



Advantages: excellent mimicking of spatiotemporal correlations.
Disadvantages: very slow to compute; subject to biases.

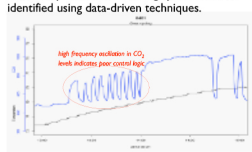




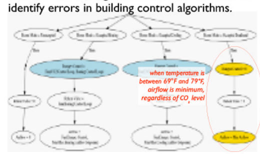
Red = missing data
Gray = reconstructed data

Insights from RSF (Research Support Facility) Building Data

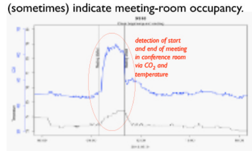
Incorrectly controlled building systems can be identified using data-driven techniques.



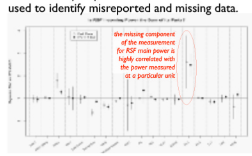
The RSF control logic codebase can be mined to identify errors in building control algorithms.



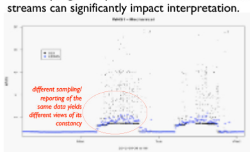
Temperature and CO_2 measurements can (sometimes) indicate meeting-room occupancy.



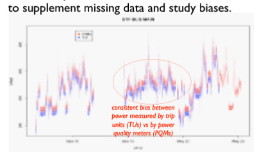
The hierarchy of power meters in the RSF can be used to identify misreported and missing data.



The sampling rates present in different RSF data streams can significantly impact interpretation.



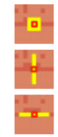
Redundancy in RSF measurements can be used to supplement missing data and study biases.



Anomaly and Error Detection in Large Spatial Datasets

Goals: Develop a reusable methodology to support:
- Processing and performing statistical analysis of very large spatiotemporal datasets
- Identification of spatial and temporal trends in raw datasets and in resulting temporal statistics
- Visualizing these phenomenon for use in analysis and in identification of anomalies
- Application of data correction and noise reduction methods

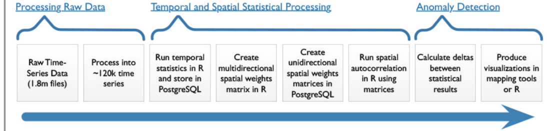
Results: Developed a reusable methodology for anomaly detection in large geospatial datasets using a combination of spatial database and spatial statistical analysis methods.

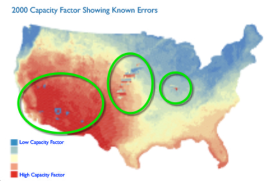


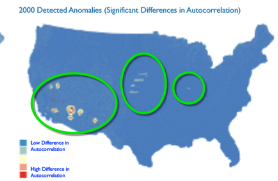
Multi-directional Autocorrelation: Statistical measure of the similarity of a feature to the features within a specified distance in all directions.

Uni-directional Autocorrelation: Statistical measure of the similarity of a feature to the features within a specified distance and within a single direction.

Anomaly Detection: Because these data represent natural phenomenon with relatively high spatial autocorrelation, differences between the horizontal, vertical, and multidirectional autocorrelation indicate anomalies or errors in the data. By calculating the largest differential between the autocorrelation methods, we can highlight and depict these anomalies in maps.







Monte Carlo Study: Imputation and Clustering Techniques

Goals: We designed a large-scale Monte Carlo simulation in order to evaluate the effectiveness of each imputation/clustering base combination (see Figure 1). In particular, we are interested in answering the following:
- Is it better to impute using multiple imputation or simple spline-based methods?
- Should we base the clustering on the autocorrelation function of each data stream or on several properties related to the wavelet representation of the stream?
- How robust is the fuzzy clustering method?

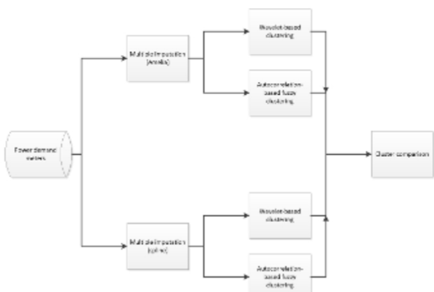
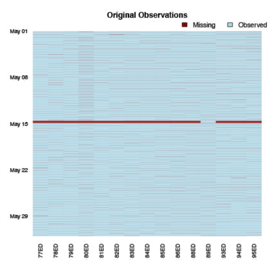


Figure 1: Analysis flow diagram illustrating the imputation and clustering approaches used in this study.

Monte Carlo Study: Data with Additional Missing Values

Example Simulation Run:
- We selected one representative month of data (May 2011) as the basis of our simulation study.
- In each simulation run, we induced additional missing values so that we could compare the methods of imputation and clustering (see Figure 2).
- The results presented in Tables 1 and 2 are based on applying the imputation/clustering combinations to the "test" data sets (right panel of Figure 2).
- We verified our results by applying these methods to the validation data set given in the left panel of Figure 2 (data not shown).



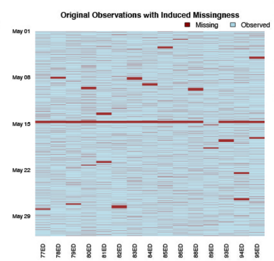


Figure 2: Missingness maps of the original, or validation, data set (left) and the same data set with more missing values, test set, added (right). Red and blue lines represent a missing and an observed value, respectively.

Monte Carlo Simulation Results and Summary

Table 1: Clustering results for a typical run in the simulation, comparing autocorrelation-function-based clustering with wavelet-based clustering. The wavelet-based clustering criteria using six clusters tended to be more stable than any of the other combinations. This is evidenced by the fact that the photovoltaic (PV) array and Data Center data streams are assigned to their own clusters.

			Five clusters				Six clusters			
Meter	Units	Description	ACF (Amelia)	ACF (Spline)	Wavelet (Amelia)	Wavelet (Spline)	ACF (Amelia)	ACF (Spline)	Wavelet (Amelia)	Wavelet (Spline)
77ED	kW	Incoming power	1	1	1	1	1	1	1	1
78ED	kW	PV Array	1	1	2	2	1	1	2	2
79ED	kW	Lighting	2	2	1	3	2	2	1	1
80ED	kWH	Electric demand	2	2	2	2	2	3	3	3
81ED	kWH	Electric demand	2	2	2	2	2	3	3	3
82ED	Factor	Water heaters	3	3	3	4	3	4	4	4
83ED	Factor	Service elevator	3	3	3	4	3	4	4	4
84ED	%	Air conditioning	4	4	4	5	4	2	5	5
85ED	%	Air conditioning	4	4	5	1	5	5	1	1
86ED	Power	Data center	3	3	3	4	6	6	6	6
88ED	Factor	Elevator	5	5	3	3	4	5	4	1
89ED	Factor	Electric demand	4	4	1	3	5	5	1	1
93ED	%	Electric demand	4	5	5	1	4	5	1	1
94ED	%	Electric demand	4	4	1	3	5	5	1	1
95ED	%	Electric demand	2	2	2	2	2	3	3	3

Table 2: Clustering results across all imputation/clustering combinations.

	Amelia		Spline	
Clusters	Wavelet	ACF	Wavelet	ACF
Five	944 (6)	890 (15)	195 (333)	88 (384)
Six	998 (3)	781 (10)	539 (153)	121 (314)

Summary: Clustering based on wavelet properties along with the multiple imputation produce the most robust results in terms of consistency of cluster membership when using six clusters. That is, 998 out of 1000 simulation trials produced the same cluster assignment (see Table 2). These results provide a strategy for imputing missing observations and clustering assignments of power demand meters.

LDRD FY13 Annual Review and Poster Session
Golden, Colorado
June 13, 2013
NREL/PO-6A20-64976